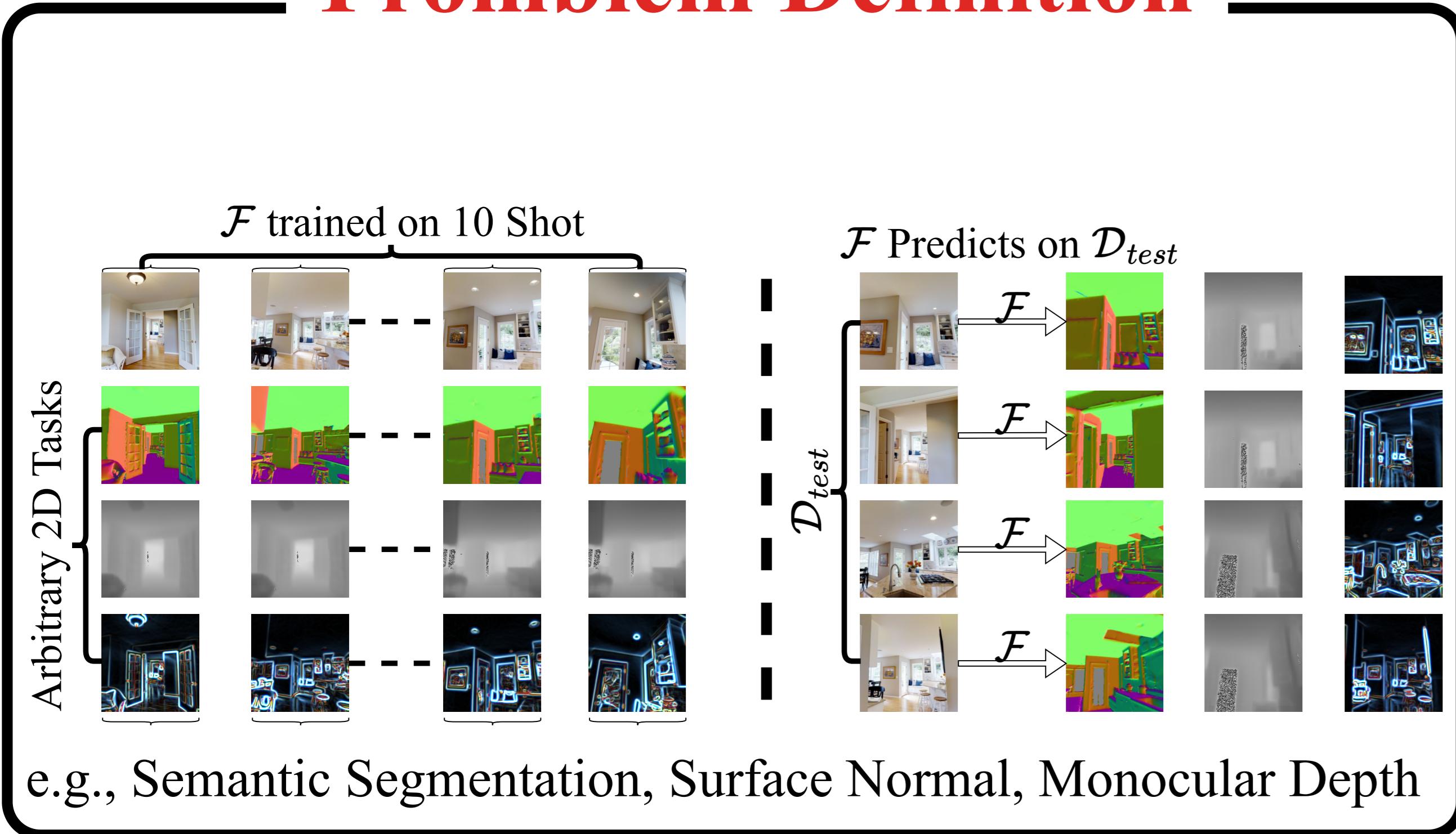


Universal Few -Shot Dense Visual Task Learner with Coupled Prompts

Jiho Choi under the supervision of Professor Hae-Gon Jeon

School of EECS, GIST, Korea

Promblem Definition



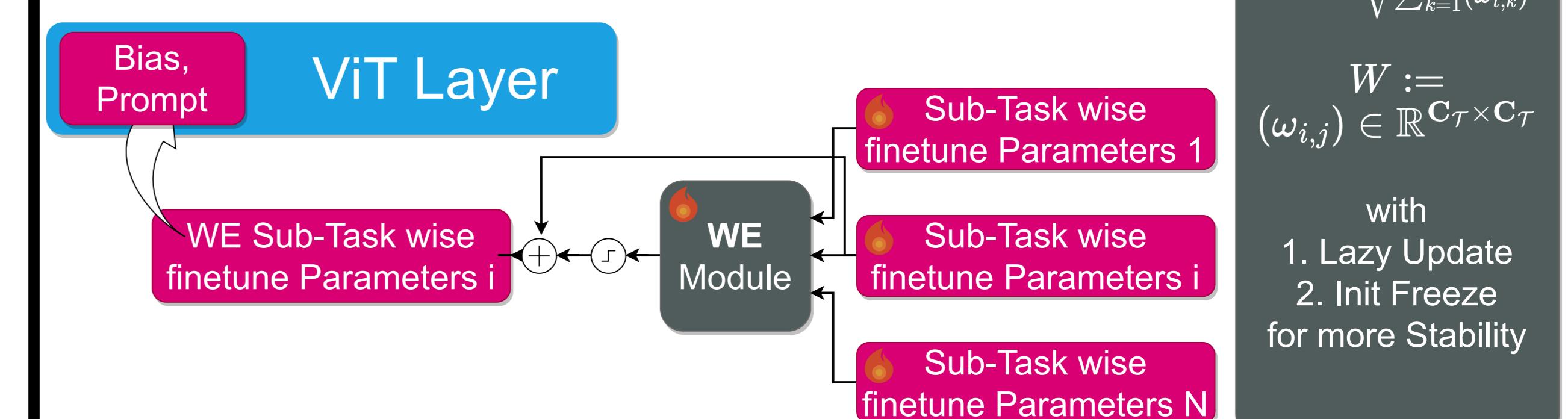
Contributions

- Devises novel method(**Weight Ensemble**) that can stabilize finetune on few-shot setup.
- Conducts extensive experiments to show effectiveness of **Prompt Tuning** on dense prediction tasks.
- Empirically shows that **knowledge** embedded in ViT's weight affects prediction performance significantly.

Methods

Weight Ensemble

- Handles Multi Channel Information
e.g., surface normal(3 channel)
- Works as Attenuator for single channel tasks
e.g., semantic segmentaion



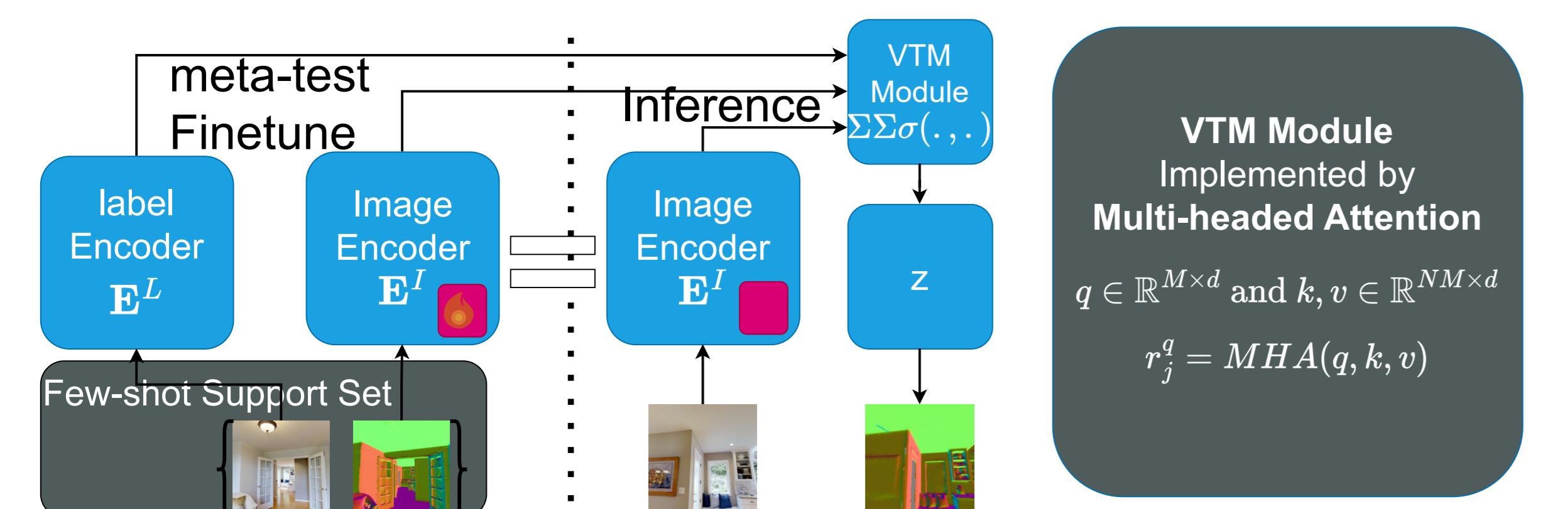
Previous Works

- VTM(ICRL 2023 oral)

$$y_j^q = \mathbf{D}^L(r_j^q), r_j^q = \sum_{i \leq N} \sum_{k \leq M} \sigma(\mathbf{E}^I(\mathbf{x}_j^q), \mathbf{E}^I(\mathbf{x}_k^i)) \mathbf{E}^L(\mathbf{y}_k^i)$$

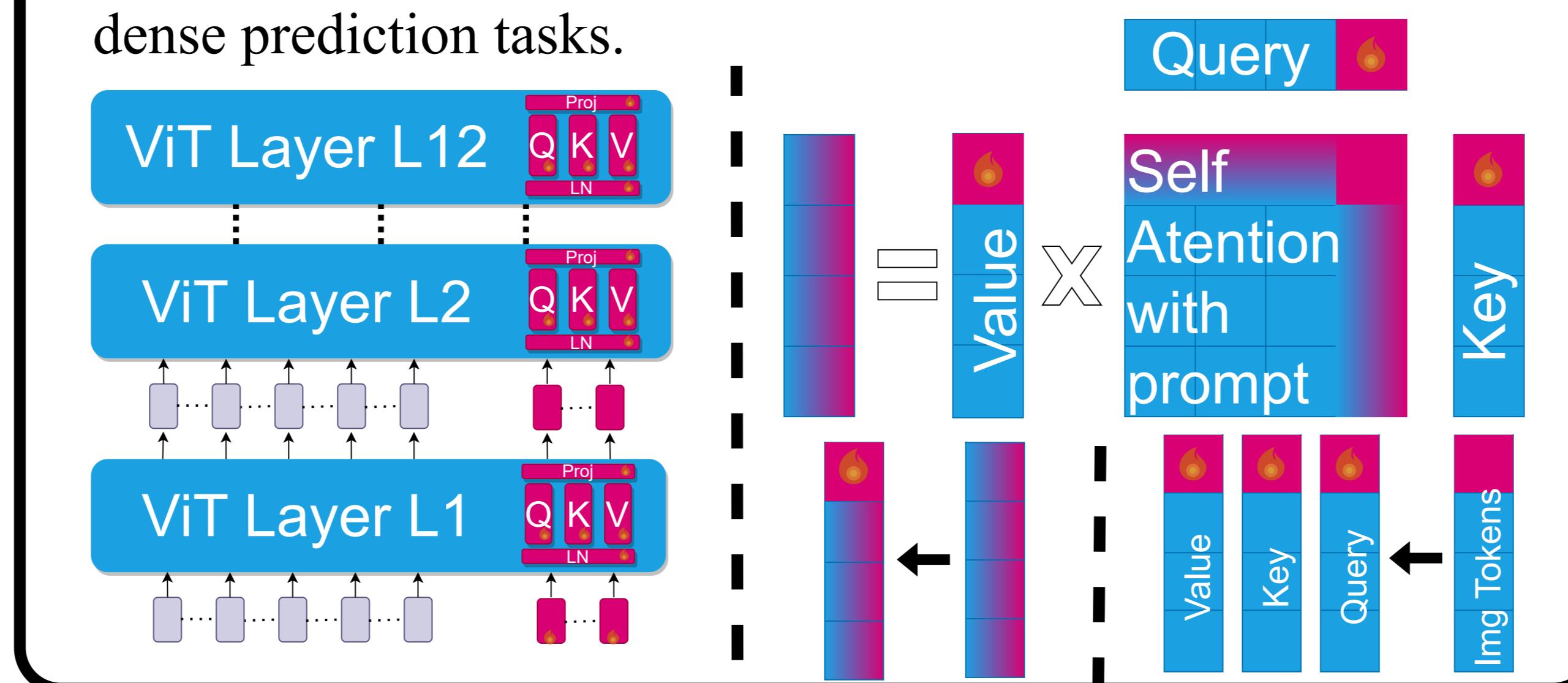
Non-parametric token matching

to achieve universal few shot learner for dense prediction tasks



Prompt Tuning

- Acts as task specific instruction that steers ViT based Encoders' pretrained knowledge.
- Shallow Prompt with Layer-wise Residual Prompts work for dense prediction tasks.



Pretrained knowledge

- Required to solve each specific task is quite different task-wisely
- Can Vision Foundation model be adopted as encoders without **catastrophic forgetting**?

ImagNet-22k
15M images
22K classes



Segment Anything(SA-1B)
1+ Billion masks
11 million images



Supervision on class prediction with cross entropy loss

Supervision on mask prediction with focal loss and dice loss

Results

- Prompt, Weight Ensemble Performances on each task

Supervision	Model	Tasks									
		Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
10-Shot	Paper Reproduce	SS	SN	ED	ZD	TE	OE	K2	K3	RS	PC
	Prompts(Ours)	0.4097	11.44	0.0741	0.0316	0.0791	0.0912	0.0639	0.0519	0.1089	0.0420
	Weight Ensemble(Ours)	0.3875	10.59	0.08998	0.0322	0.08729	0.09897	0.06462	0.04853	0.1096	0.04229

- Performances on each task for different pretrained weight

Supervision	Model	Tasks									
		Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
10-Shot	Paper Reproduce	SS	SN	ED	ZD	TE	OE	K2	K3	RS	PC
	scratch	0.4097	11.44	0.0741	0.0316	0.0791	0.0912	0.0639	0.0519	0.1089	0.0420
	IN-22k	0.3875	10.59	0.08998	0.0322	0.08729	0.09897	0.06462	0.04853	0.1096	0.04229

Mono Depth; MSE 0.3222(bias), 0.0314(Prompt)



Surface Normal; mErr 10.59(bias), 10.14(Prompt)

